基于知识图谱的“一带一路”投资问答系统构建*

■ 陈璟浩[1] 曾桢[2] 李纲[3]
[1] 广西大学公共管理学院 南宁 530004 [2] 贵州财经大学信息学院 贵阳 550025
[3] 武汉大学信息资源中心 武汉 430072

摘 要：[目的/意义]基于知识图谱的“一带一路”投资问答系统有效整合多种来源的信息资源，能为用户提供快捷、准确、高质的“一带一路”投资信息，具有重要的研究和应用意义。[方法/过程]对“一带一路”投资相关的信息进行采集、处理与整合，在专家指导下构建“一带一路”投资知识图谱。在此基础上，问答系统的各部分功能得以实现，包括：用户问题预处理、问题分类、问题模板匹配及答案查询。[结果/结论]实验结果表明，该系统能有效回答“一带一路”投资相关问题。

关键词：问答系统 知识图谱 一带一路 系统构建
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.12.011

1 引言

“一带一路”是新时期中国形成全方位开放格局的重要路径[1]。自2013年，习近平总书记提出“一带一路”倡议以来，我国企业便开始积极向“一带一路”沿线国家投资，2013－2018年，直接投资总额超过900亿美元，与沿线国家新签对外承包工程合同额超过6 000亿美元[2]。随着投资活动增加，企业对东道国国情、投资环境、投资政策、投资手续等信息需求也随之增多。但从现有情况看，一方面，仅依靠互联网搜索引擎来获取“一带一路”投资相关信息，存在信息数量大、信息冗余、信息质量良莠不齐等问题，需要耗费大量人工才能获取其中的知识；另一方面，“一带一路”投资信息资源的多源异质、结构松散等特点，使其整合性和关联性差，难以提供规范的数据和实现丰富的语义表达。为此，如何对网络中“一带一路”投资相关信息资源进行梳理，如何提升各种信息资源的利用率，为用户提供准确信息并减少相应的查询时间，便成为当前亟待解决的问题。

自动问答系统能接受人们提出的自然语言问题，在知识库中查找相应答案，并返回给用户[3]。与传统搜索引擎相比，自动问答系统增强了用户获取知识的便捷性，节省了信息筛选时间，也提高了信息质量。传统的自动问答系统大多基于文档检索，使用关键词或模板匹配的方式查询答案，而答案的数据来源基本都是非结构化的文本，在查询精度、问题推理、语义关联方面先天不足。知识图谱的出现，在一定程度改变了这种情况。知识图谱是以图的形式表现客观世界中的实体（概念、人、事物）及其之间关系的知识库[4]，知识图谱以三元组作为表示形式。将知识图谱技术运用于自动问答系统有助于从海量文本信息中抽取结构化的知识，将不同来源数据进行融合，形成富含语义关系的知识网络，可以为问答系统提供高质量的信息。通过集成知识图谱，问答系统的数据精度、数据关联性、数据结构化水平得到显著提升，增强了问题语义和知识语义的理解和匹配。基于此，构建一个基于知识图谱的“一带一路”投资问答系统，在一定程度上能解决前述信息获取过程中所出现的问题。

综上，笔者试图提出一套基于知识图谱的“一带一路”投资问答系统设计实现方案。文章首先介绍了国内外问答系统的研究现状，并进行简要评述；然后对所构建系统的设计思路和功能架构进行介绍；接着阐述

* 本文系国家自然科学基金重大项目“国家安全大数据综合信息集成与分析方法”（项目编号：71790612）和国家自然科学基金项目“基于数据挖掘的跨区域网络情报智能分析研究——以东盟十国为例”（项目编号：71663005）研究成果之一。
作者简介：陈璟浩（ORCID：0000－0001－8768－8562），讲师，博士；曾桢（ORCID：0000－0001－8481－3567），副教授，博士；李纲（ORCID：0000－0001－5573－6400），主任，教授，通讯作者，E-mail：jhchen114@qq.com。
收稿日期：2020－01－04 修回日期：2020－04－12 本文起止页码：95－105 本文责任编辑：徐健

了系统中关键技术的实现过程;最后对构建的系统进行实验以证明系统的可用性,并对未来工作进行展望。

2 相关研究回顾

依据回答范围不同,自动问答系统可以分为开放域问答系统和限定域问答系统两类。开放域问答系统的问答并不受具体领域的限制,可以对多个领域的提问进行回答,通常其会利用 Web 数据资源冗余的特点,通过统计方法来查找正确答案[5]。另外,开放域问答系统用户提出的问题相对简单,用词也是一些日常用语,对用户提问范围一般没有限制,其答案主要来自 Web 资源[6]。开放域问答系统通常会使用一些通用语义资源,如 WordNet、HowNet、常识图谱 CYC 等,以及基于语义网技术的关联数据(linked data),如 FreeBase、Dbpedia 等[7]。目前,比较有代表性的开放域问答系统有:英文问答式检索系统 Ask Jeeves、麻省理工学院开发的 START[8]、多语系自动问答系统 AnswerBus[9] 及 IBM 的 Watson[10] 系统。

限定域问答系统一般只能处理限定领域的相关问题,相对于开放域问答系统,它处理的问题要更专业也更为复杂。其面向的对象更多是熟悉此领域的用户(如领域专家),他们一般会使用一些领域术语来查询,对反馈的答案质量要求也比较高。限定领域问答系统通常以具体的目标和任务为导向,这就决定了其需要领域知识库、领域词典等作为支持,这在某些程度上也决定了系统所能回答问题的范围。大部分限定领域问答系统因为领域较窄,用户量较小等原因,获取和建设高质量的语料资源显得尤为宝贵和重要。限定域问答系统经历了长期发展,从 20 世纪 60 年代基于结构化数据的问答系统,如 Baseball[11] 和 Lunar[12],到 70 年代、80 年代基于计算语言学的问答系统,如 Berkeley Unix Consultant[13],到 90 年代基于自由文本的问答系统,再到本世纪初出现的基于常问问题数据 FAQ 的问答系统,其研究成果层出不穷,技术有了长足进步。自 2012 年谷歌公司推出基于知识图谱技术的搜索产品以来,该技术在人工智能研究领域便得到了广泛的应用,基于知识图谱的限定域问答系统研究也成为主流。

当前,生命科学、生物医学、图书情报学等诸多领域都开展了基于知识图谱的自动问答系统研究。M. Vargas-Vera 等开发了一款名为 AQUA 的学术领域问答系统,知识图谱技术在系统中被用于查询细化、问题推理和相似度计算[14];A. Ben-Abacha 等结合医学领域知识、自然语言处理技术和知识图谱技术开发了 MEANS 医疗问答系统[15];A. H. Asiaee 等开发了名为 OntoNLQA 的生物医学领域问答系统,该系统由自然语言处理、实体识别、图谱匹配、语义关联和答案检索 5 个主要部分构成[16];X. Xie 等构建了《自然语言处理》课程自动问答系统,该系统包含 4 个处理模块,基于图谱的知识库、问题分析模块、答案抽取模块和标准答案扩展模块[17];A. Abdi 等建立了一个物理领域问答系统,该系统采用了一种基于语义和句法信息的推理映射方法,将用户提问转化为知识库查询语言[18];A. Agarwal 等构建了一个融合教育语义的动态概念网络模型,该模型提升了教育领域问答系统 EDUQA 的准确率[19];马晨浩创建了甲状腺知识图谱,并在此基础上设计实现了面向甲状腺诊疗的自动问答系统[20];曹明宇等构建了原发性肝癌知识图谱,实现了流水线式的问答系统[21];杜泽宇等提出了一套流式的中文知识图谱自动问答系统 CEQA,能够较好地完成电商领域商品咨询以及统计推理等复杂问题[22];陆伟等根据武汉大学图书馆的业务需求构建了图书馆领域自动问答系统,该系统引入了知识图谱技术,并建立了多源数据融合知识库[23]。上述研究从多方面阐述了基于知识图谱的自动问答系统构建过程,对本研究有借鉴意义。

总的来看,作为语义网的支撑,知识图谱在自动问答领域起着至关重要的作用,其已成为组织、表达、管理海量、异构、动态数据的有效方式。“一带一路”倡议作为我国的一项重要发展战略,受到政府部门及研究机构的重视,建设了许多网站、平台和数据库,如中国“一带一路”网、“一带一路”频道、“一带一路”数据库等,这些资源对指导企业投资“一带一路”沿线国家具有重要价值。然而,从目前情况来看,这些资源利用率并不高,主要集中在文本层次利用,并未深入到内容层面,服务目标单一,相关信息资源整合力度不够。因此,通过将大量有关“一带一路”投资相关信息进行汇总,创建“一带一路”投资知识图谱,并在此基础上设计“一带一路”投资自动问答系统,能帮助用户快速、准确、充分地了解相关知识,并填补当前研究空白。

3 系统框架

实现基于知识图谱的“一带一路”投资问答系统,首选需要解决的是采集和获取用于支持问答的数据资源,然后对数据资源进行处理和组织,形成问答语料数据;在此基础上建立知识图谱并构建知识库;知识库建好后,进一步对用户输入的问题进行分析处理、匹配查询,获取最终答案。依据此思路,笔者实现的自动问答

系统可分为3大模块:数据获取与处理模块、知识图谱构建模块以及问题分析与答案获取模块,系统框架如图1所示:

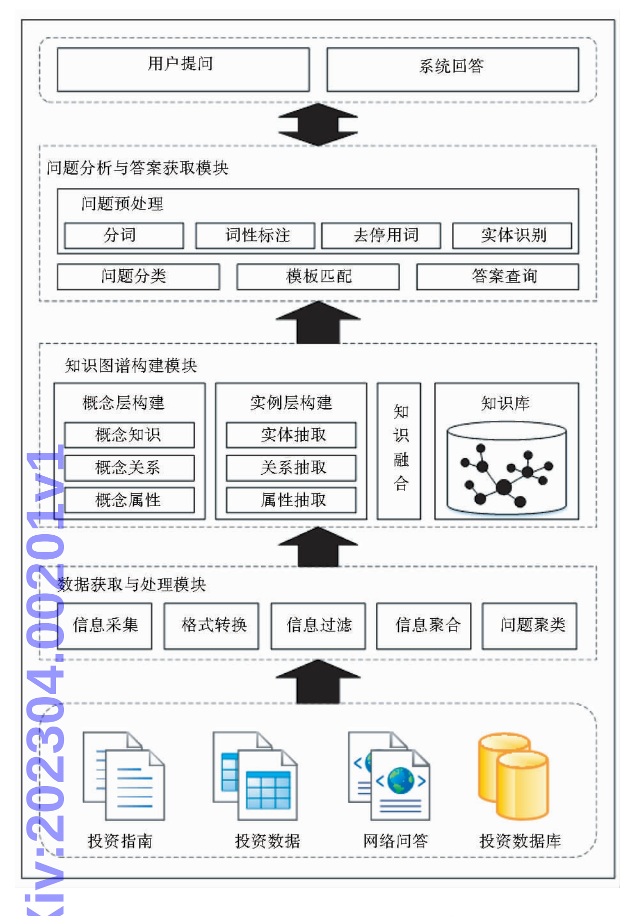


图1 "一带一路"投资问答系统框架

(1)数据获取与处理模块。数据获取与处理模块包含的功能有信息采集、格式转换、信息过滤、信息聚合和问题聚类。信息采集是通过网络爬虫和人工方式从各种数据源采集和下载"一带一路"投资相关信息,如百度知道、知乎问答平台中有关"一带一路"沿线国家投资的问答数据、商务部《中国对外投资企业名录》、商务部《外商投资企业名录》、中国"一带一路"网中的《"一带一路"沿线国家投资指南》等;格式转换是将采集和下载的电子表格、PDF文档转换为数据库中的数据表;信息过滤是对采集的冗余信息、噪声信息、无关信息进行过滤;信息聚合是将不同来源的信息进行整合,如《中国对外投资企业名录》里只有对外投资企业的名称,没有企业所属行业、所属地区等属性信息,这时可利用"企查查"网的企业信息数据库对《中国对外投资企业名录》中的企业信息进行补全和整合;问题聚类是对百度知道、知乎问答平台的问题进行聚类,梳理和归纳网民在这些问答平台上提出的"一带一路"投资相关问题,为后续划分问题类别、建立知识图谱和问题模板提供参考。

(2)知识图谱构建模块。知识图谱构建模块采用自上而下的方法对图谱进行构建,包含的工作有图谱概念层构建、实例层构建、知识融合及生成知识库。概念层构建是对知识图谱的"骨架"进行搭建,其将对"一带一路"投资知识图谱涉及的概念、术语、关系和属性进行定义,明确图谱的范围,规范图谱的表达。概念层存储的是经过提炼的知识,通常采用本体库来管理,借助本体库对公理、规制和约束条件的支持能力来规范实体、关系以及实体的类型和属性等对象之间的联系[24]。实例层构建是在概念层基础上展开,对实体、关系和属性的抽取工作,其中实体抽取又称之为命名实体识别,本文采用基于规则和词典的方法;关系抽取负责提取实体间的关联关系形成知识网络,采用基于词典驱动的方法;属性抽取是从不同数据源汇集实体的属性信息,实现对实体的完整勾画,抽取方式与关系抽取相同。实例抽取完毕后,还需采用知识融合方法,对抽取结果进行组织,以消除矛盾和歧义,具体技术包括:实体链接、知识合并等。在完成知识融合相关工作后,事实便以"实体－关系－实体"或"实体－属性－属性值"的三元组形式存储,形成一个图状知识库。

(3)问题分析与答案获取模块。问题分析与答案获取模块包含的功能有问题预处理、问题分类、模板匹配和答案查询。问题预处理功能是对交互界面中用户输入的自然语言问题进行处理,包括分词、词性标注、去停用词和实体识别。问题分类是依据数据获取与处理阶段划分的问题类别,利用文本自动分类技术,将处理好的用户提问划分到相应类别中去,这能有效减少候选答案的空间,提高系统返回正确答案的概率。问题分类完成后需要对问题进行理解,本文采用基于模板匹配的方法[25]。问题模板根据问句类别中的常见问题设计,其作用是将用户提问映射为相应的数据库查询语言。模板匹配过程是通过相似度算法计算用户提问与预先准备好的问句模板之间的相似度值,当相似度值超过某一阈值,则认为匹配成功。另外,当出现多个模板相似度值超过阈值时,则使用相似度值最高的模板。模板匹配完毕后,根据识别出的实体名及关系类型,理解问题语义,在构建好的"一带一路"投资知识图谱中查询对应的实体或属性,将查询结果生成符合对话逻辑且语法通顺的答案返回给用户。

4 系统主要实现过程

4.1 数据获取与处理

4.1.1 问答数据获取

在构建“一带一路”投资问答系统之前,首先需要收集领域知识,笔者主要采用网络爬虫和人工下载方式。其中,网络爬虫模块是利用 HTML 解析器 Jsoup,集成 HTTPClient 编程工具包,通过正则表达式对页面中的数据进行采集。采集内容包括:商务部《中国对外投资企业名录》(http://femhzs.mofcom.gov.cn/fecpmvc/pages/fem/CorpJWList.html)数据 25 034 条,商务部《外商投资企业名录》(http://www.fdi.gov.cn/1800000121_10000207_8.html)数据 3 180 条。问答数据则是通过构建“国家名 + 投资”的关键词检索式,如“新加坡 + 投资”,对百度知道和知乎网站中与“一带一路”沿线国家投资相关的问答对进行爬取,共计 31 555对。另外,为了保证知识的权威性,笔者还通过人工方式对中国“一带一路”网中的《“一带一路”沿线国家投资指南》(https://www.yidaiyilu.gov.cn/info/iList.jsp? cat_id = 10148)进行下载。需要说明的是,本系统所涉及的“一带一路”沿线国家共 64 个,依据北京大学“一带一路”数据分析平台的界定方案划分(见表1)[26]。上述获取的数据及下载的投资指南覆盖了这 64 个国家。最终,本系统构建的“一带一路”投资相关三元组数量为 39 982 对。

表 1 “一带一路”沿线 64 个国家及区域分布

区域	数量(个)	国家名称
东南亚	11	新加坡、印度尼西亚、马来西亚、泰国、越南、菲律宾、柬埔寨、缅甸、老挝、文莱、东帝汶
南亚	7	印度、巴基斯坦、斯里兰卡、孟加拉国、尼泊尔、马尔代夫、不丹
西亚、北非	18	阿联酋、阿塞拜疆、科威特、土耳其、卡塔尔、阿曼、黎巴嫩、沙特阿拉伯、巴林、以色列、也门共和国、埃及、伊朗、约旦、叙利亚、伊拉克、阿富汗、巴勒斯坦
中东欧	22	波兰、俄罗斯、阿尔巴尼亚、格鲁吉亚、爱沙尼亚、立陶宛、亚美尼亚、斯洛文尼亚、保加利亚、捷克共和国、匈牙利、马其顿、塞尔维亚、罗马尼亚、乌克兰、斯洛伐克、克罗地亚、摩尔多瓦、白俄罗斯、拉脱维亚、波黑、黑山
中亚及蒙古	6	哈萨克斯坦、吉尔吉斯斯坦、土库曼斯坦、塔吉克斯坦、乌兹别克斯坦、蒙古

4.1.2 问答数据处理

在获取问答数据后,还需要对数据进行处理,包括:将各国的投资指南 PDF 转换为文本文档进行存储(主要通过 PDFBox 开源软件包进行处理);将《中国对外投资企业名录》数据和《外商投资企业名录》数据整合,形成“一带一路”企业投资数据库,并利用网络爬虫采集“企查查”网站数据,对数据库中的企业属性信息进行补全,如企业所在地区、行业、类型、地址、经营范围等。另外,作为后续知识图谱和问题模板构建的参考,还需对采集的百度知道和知乎问答数据进行过滤和聚类。过滤主要是剔除与“一带一路”投资主题无关的问题、对重复的问题去重以及删除空值数据。聚类则是利用聚类算法对问题聚类,考虑到聚类类别数的自动划分以及传统空间向量模型特征高维稀疏、对语义关系缺乏考虑的问题,笔者采用自动划分聚类数的聚类算法 DBSCAN[27],特征提取方法采用 Word2Vec 结合 TF-IDF 进行文本表示[28],其中 DBSCAN 算法的 eps 值取值为 0.5,训练 Word2Vec 模型的语料是维基百科中文语料加上百度知道和知乎上采集的“一带一路”投资问答语料共 1.2GB,采用 Skip-Gram 模型训练,单词维数 300,训练窗口 10。自动聚类得到 2 240 个问题类,通过人工审核、筛选归并,最终保留问题类 83 个,共 10 602 个问答对。另外,为了确保所有问题回答的准确性,对于百度知道和知乎问题的回答数据,项目组还招募了 5 名研究生对问题回答进行审核,为了确保答案的准确性,学生 2 人一组,对答案进行筛选。首先如果点赞数最多,回答时间与系统构建时间最接近,则作为准确答案;如遇到点赞数不高,但回答时间与系统构建时间最接近的情况,由学生对候选答案进行比对(点赞数最多的和回答时间最接近的答案比对),最后 3 人投票,票数最高的答案作为候选答案。

4.2 知识图谱构建

本文的知识图谱采用自顶向下的方法构建,依次为图谱概念层构建和实例层构建,概念层主要是结合前述整合的各类数据,对知识图谱涉及的术语、概念及关系展开抽取和定义,明确图谱的整体范围,实例层则是在概念层的约束下填充数据,最终形成结构化的知识图谱。以下将给出“一带一路”投资概念层、实例层的具体构建方法和存储方式。

4.2.1 概念层构建

“一带一路”投资知识图谱的概念层设计是在领域专家的帮助下,结合《“一带一路”沿线国家投资指南》、百度知道、知乎问答、《中国对外投资企业名录》等相关知识构建而成。为了满足用户的提问需求,笔者构建了“一带一路”沿线国家投资图谱概念层及“一带一路”企业投资图谱概念层,主要工作包括:领域概念归纳和领域关系及约束定义。

(1)领域概念归纳。“一带一路”沿线国家投资图谱概念层包括：国家基本信息、投资基本信息、投资法规政策、投资手续办理、投资注意事项、遭遇困难求助 6 个核心概念。其中，国家基本信息是指被投资国的国家概况，包括：国家历史、政治环境、地理环境、社会文化 4 个子概念；投资基本信息主要反映被投资国的投资潜力，包括：经济表现、国内市场、基础设施、对外经贸、金融环境、证券市场、商务成本 7 个子概念；投资法规政策是指被投资国家的相关投资法规政策，包括：对外贸易法规政策、外国投资市场准入、企业税收规定、外国投资优惠、特殊经济区、劳动就业规定、外企土地投资、外企证券交易、环境保护法律、反商业贿赂法规、外企承包工程、知识产权法规、投资合作法律、商务纠纷 14 个子概念；投资手续办理主要反映被投资国投资手续的办理方法及流程，包括：投资注册企业、承揽工程程序、申请专利、注册商标、报税手续、工作证办理、投资咨询机构 7 个子概念；投资注意事项是指在被投资国进行投资时需要注意的情况，包括：贸易注意事项、承包工程注意事项、劳务合作注意事项、需防范的风险和其他注意事项 5 个子概念；遭遇困难求助则是指在被投资国遭遇困难后的求助方式，包括：寻求法律保护、寻求政府帮助、应急预案和中国驻当地使馆保护 4 个子概念。此外，为了明确问题，使概念更加具体，在一些 3 级子概念下，还进一步划分了 4 级子概念，如社会文化 3 级子概念就进一步划分了民族、语言、宗教、习俗等 4 级子概念。最终，“一带一路”沿线国家投资图谱概念层共构建了 6 个核心概念，41 个 3 级子概念和 97 个 4 级子概念。“一带一路”沿线国家投资图谱如图 2 所示：

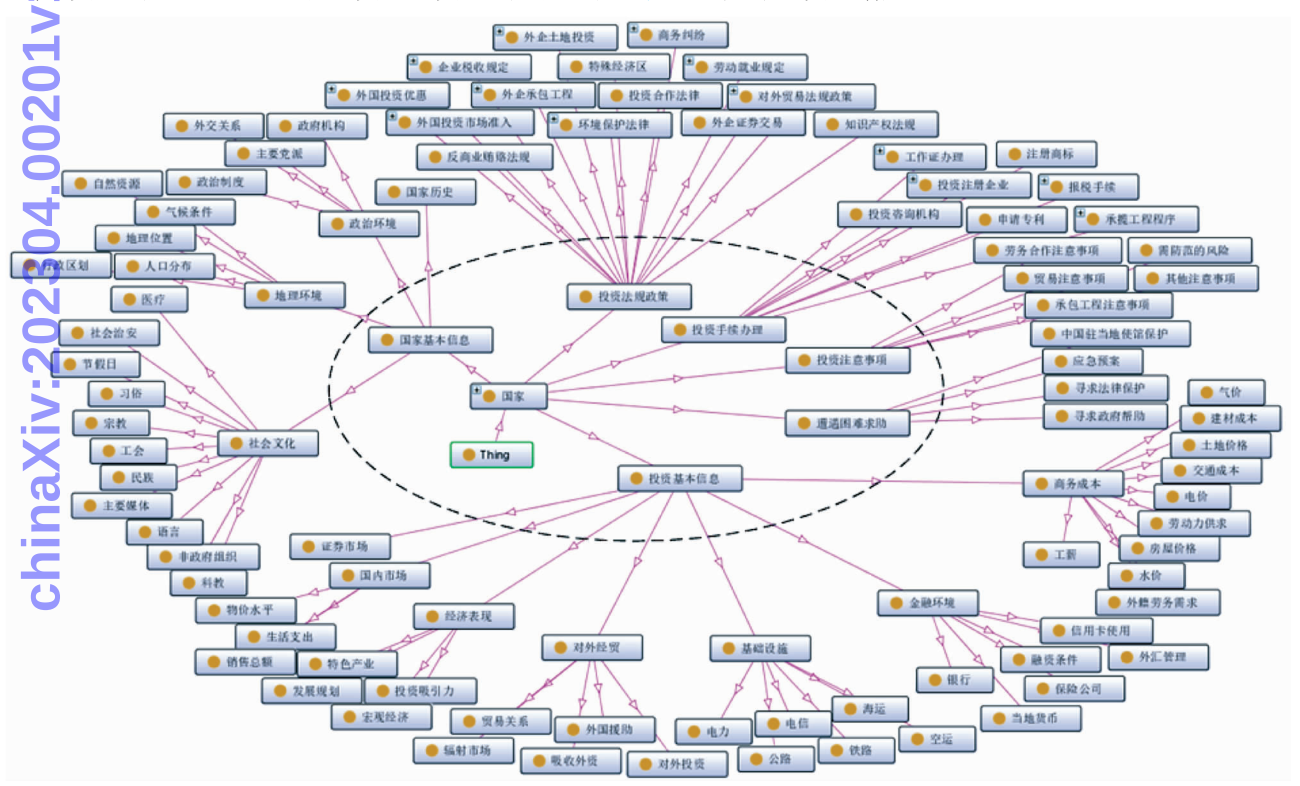


图 2　“一带一路”沿线国家投资图谱

相对于沿线国家投资图谱，“一带一路”企业投资图谱相对简单，主要服务于“一带一路”企业投资情况问答，其包括：投资国、所属国、行业、类型、地址、注册资本、实缴资本、经营范围 8 个 2 级子概念，及所属地区和投资地区 2 个 3 级子概念。“一带一路”企业投资图谱如图 3 所示：

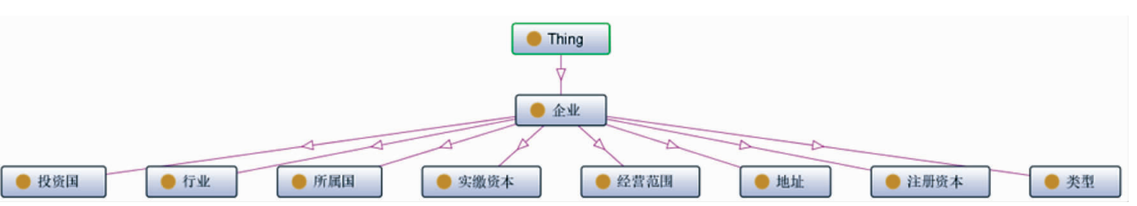


图 3　“一带一路”企业投资图谱

（2）领域关系及约束定义。关系是概念层的核心基本要素，它描述了领域中的概念和实例之间的作用关系，决定了知识图谱的丰富程度。笔者主要通过 2 种方法来定义概念之间的关系，一是参考质量较高的数据源，如《“一带一路”沿线国家投资指南》；二是从关系数据库表中抽取现成的关系模式。如《中国对外投资企业名录》。最终确定了 6 大类关系，如表 2 所示：

表 2 概念层图谱实体关系类别

关系类别	含义	举例
同义关系	相同或相似表达	<印度尼西亚，同义词，印尼>
属性关系	实体的属性和属性值	<恩盛电器，注册资本，9000 万>
层次关系	根据范围确定概念及子概念	<国家基本信息，包含，社会文化>
整体－部分关系	实体由整体和部分构成	<电信，构成，基础设施>
投资关系	实体的投资行为	<美珍香，投资，中国>
隶属关系	实体的隶属关系	<美珍香，属于，新加坡>

4.2.2 实例层构建

概念层构建完成后，便可在其基础上构建实例层。实例层构建任务主要是从前述获取和处理好的记录中抽取与概念层相匹配的“一带一路”投资知识。这一过程既要对结构化数据进行处理也要对半结构化和非结构化数据进行处理。实例层构建的目标就是从不同来源的记录中提取“一带一路”投资实体及关系，并将其表示为三元组的形式。具体来说，实例层构建的工作包括：实体抽取、关系抽取和属性抽取。

（1）实体抽取。根据概念层确定的概念，从数据记录中抽取相应的“一带一路”国家名称、企业名称等，构建相应的实体节点，形成概念到实体间的映射。如泰国、越南、新加坡等国家实体，以及美珍香、亚太纸业、中巨财富等企业实体。

（2）关系抽取。根据概念层确定的关系构建实体间的关系，并根据概念间的关系名称确定实体间的关系名称。如：在概念层企业和国家之间存在投资关系，依据《中国对外投资企业名录》，企业万达国贸集团与国家新加坡存在投资关系，则在这两个实例之间添加投资关系。同时。国家基本信息概念包含社会文化子概念，则具体到实例，新加坡国家基本信息与新加坡社会文化就是包含关系。另外，在概念层中定义了同义关系，则将具有同义关系的实体与其别称进行关联。

（3）属性抽取。“一带一路”投资知识图谱的属性抽取主要是依据实体所对应的概念层含有的属性，抽取属性值。《“一带一路”沿线国家投资指南》作为高质量的数据源，属性抽取即可直接借助这些资料，获取实体的属性及属性值。如：新加坡投资指南中就具体介绍了新加坡的习俗，因此，这部分内容就可直接作为新加坡习俗的属性值。又如：在“一带一路”企业投资数据库中，美的集团具有属性“行业”，其属性值为“制造业”，则可构建<美的集团，行业，制造业>的“实体－属性－属性值”三元组。表 3 给出了概念层到实例层的部分映射：

表 3 “一带一路”沿线国家投资图谱概念层到实例层的映射（部分）

概念层		实例层
国家	国名	印度尼西亚
	别名	印尼
国家基本信息		印度尼西亚基本信息包括：国家历史、社会文化、地理环境、政治环境 4 个方面
政治环境		印度尼西亚政治环境包括：政治制度、主要党派、政府机构、外交关系 4 个方面
政治制度		【政治制度】实行总统制，总统既是国家元首，也是政府首脑，同时掌管三军。总统、副总统均由全民直选产生，任期 5 年，总统可连任一次。现任总统佐科・维多多，2014 年 10 月通过直选担任新一届总统，副总统为优素夫・卡拉，任期至 2019 年。本届内阁于 2014 年 10 月组建，2015 年 8 月改 组，2016 年 7 月再次改组。现任阁员 34 人，任期至 2019 年……
……		……

4.2.3 知识存储

笔者采用当前比较流行的开源图数据库 Neo4j 进行知识图谱的存储。Neo4j 用 java 语言实现，以网络的方式对结构化数据进行存储，与关系数据库相比，Neo4j 能很好地解决数据价值密度低、数据量大的问题，其提供了完善的图查询语言，支持各种图挖掘算法。Neo4j 提供 Cypher 语句来导入和查询数据。对于大规模数据，Neo4j 还提供了 neo4j-import 工具，可以快速地将大量实体和关系导入图数据库。笔者将构建的“一带一路”投资相关三元组通过 Cypher CREATE 语句、Cypher LOAD CSV 语句以及 neo4j-import 工具导入 Neo4j 数据库。图 4 展示了 Neo4j 数据库中的“一带一路”投资知识图谱的部分三元组关系。

4.3 问题分析与答案获取

完成“一带一路”知识图谱的构建和存储，便可开展问题分析与答案获取方面的工作。此部分的内容有：对用户输入的自然语言问题进行预处理，对问题进行分类，对问题进行模板匹配得到计算机查询语句，在知识图谱中展开答案查询。

4.3.1 问题预处理

问题预处理主要是对用户提出的自然语言问题进行分词、词性标注、去停用词和实体识别。笔者主要采

陈璟浩，曾桢，李纲. 基于知识图谱的“一带一路”投资问答系统构建[J]. 图书情报工作,2020,64(12):95－105.

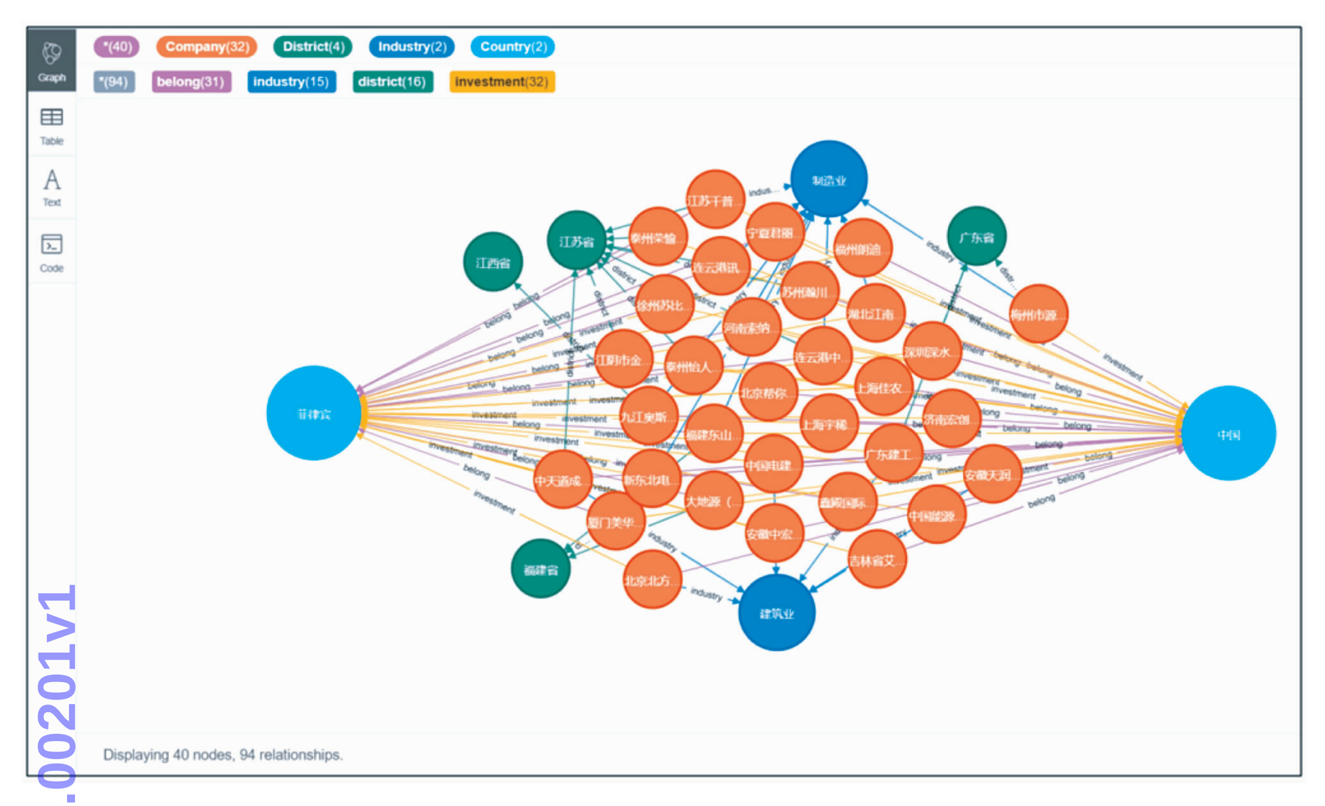


图 4 “一带一路”投资知识图谱部分三元组关系

用 HanLP 开源软件包(https://github.com/hankcs/HanLP)进行分词,该软件包含条件随机场 CRF 分词,词性标注和命名实体识别功能。同时,还构建了一个由国家别名和“一带一路”企业名录组成的自定义词典,以提高分词和命名实体识别的准确性。去停用词主要是在分词后剔除没什么实际意义的词语,包括:语气助词、副词、介词、连接词,如:的、在、啊等。用户输入的自然语言问题在经过问题预处理以后,在系统后台将被抽象化,以备后续分类和模板匹配。表 4 是问题预处理的输入输出结果范例:

表 4 问题预处理的输入输出结果范例

用户输入的自然语言问句	经过问题预处理后的问句
新加坡是个什么样的国家?	[$ country1] 是 个 什么样 国家
在新加坡投资的中国企业有哪些?	[$ country1] 投资 [$ country2] 企业 有 哪些
美珍香在哪些国家进行了投资	[$ enterprise1] 哪些 国家 进行 投资

4.3.2 问题分类

经过问题预处理后,需要将预处理的结果自动划分到预先定义好的问题类别中去。在自动问答系统中对问题进行分类,一方面可以有效地减少候选答案的空间,提高系统返回正确答案的准确性;另一方面,通过问题分类,可以根据不同的问题采取不同的答案选择策略。笔者结合 4.1.2 节整理好的百度和知乎问答及《“一带一路”沿线国家投资指南》,将“一带一路”投资领域常问问题分为 6 个类别,分别是事实类问题、方法类问题、列表类问题、计数型问题、判断类问题和其他类问题。事实类问题主要回答“是什么”;方法类问题主要回答“怎么办”;列表类问题主要是对数据库进行查询;计数型问题是对符合提问条件的数据进行统计;判断类问题主要是肯定或否定回答;其他类问题是不属于上述 5 类的问题,每类问题都有比较典型的特征词,对分类有一定指导意义,详细分类如表 5 所示:

表 5 “一带一路”投资相关问题分类

序号	问句类型	范例	抽象化范例
1	事实类问题	新加坡的历史。	[$ country1] 历史
2	方法类问题	在新加坡注册企业如何办理相关手续?	[$ country1] 注册 企业 如何 办理 相关 手续
3	列表类问题	投资新加坡的中国企业有哪些?	投资 [$ country1] [$ country2] 企业 有 哪些
4	计数型问题	投资新加坡的中国企业有多少?	投资 [$ country1] [$ country2] 企业 有 多少
5	判断类问题	在新加坡投资房产好不好?	[$ country1] 投资 房产 好不好
6	其他问题	在新加坡投资应该注意什么?	[$ country1] 投资 应该 注意 什么

问题类别确定后,笔者将各类别下的问题转换成抽象化范例并对其进行去重(见表5),最终参与文本自动分类的抽象化范例为1 853个。自动分类算法为SVM,特征提取算法与问题聚类中所用的算法一样,采用Word2Vec结合TF-IDF进行文本表示[28]。利用准确率、召回率和F-测度值对自动分类结果进行测评,测评结果如表6所示:

表6 文本自动分类的准确率、召回率和F-测度值

名称	准确率	召回率	F-测度值
公式	$\rho = \frac{a}{a+b} \times 100\%$	$r = \frac{a}{a+c} \times 100\%$	$\text{F-Measure} = \frac{2\rho r}{\rho + r}$

公式中a为分类正确的样本数据,b为错误的划分到该类别的样本数据,c为属于该类但未被区分出来的样本数据,ρ为准确率,r为召回率,*F-Measure*为F-测度值。自动分类结果如表7所示:

表7 问题自动分类结果

类名	准确率	召回率	F-测度值
事实类问题	88.66	89.20	88.93
方法类问题	91.59	92.37	91.98
列表类问题	91.24	87.47	89.32
计数型问题	96.57	96.51	96.53
判断类问题	92.56	93.22	92.89
其他问题	87.86	88.34	88.10

分类结果显示,本文采用的文本自动分类算法获得的最高F-测度值为96.53%,平均F-测度值为91.29%,具有一定效果,能满足实际应用需要。

4.3.3 问题模板匹配

在对问题进行分类后,则要将用户输入的问题转化为对应的模板,以便后续答案查询。具体来说,问题模板匹配流程包括:①针对常见问题设置相应的模板集;②将用户输入的自然语言问题抽象化并与模板集进行相似度匹配,选择相似度最高的模板。

(1)问题模板设置。根据问题中包含的实体数量和实体类别,针对每一种类型和每一种情况设计了一个包含6个类别、2个层次的具有一定冗余性的问题模板集。其中,6个类别为4.3.2节问题分类中划分的问句类型,2个层次为主层次和附属层次,主层次模板直接与图数据库Neo4j的Cypher查询语句对应,其在用户提问中出现频次最多;附属层次模板与主层次模板关联,其代表的语义意义与主层次模板的语义意义一致,其建立的目的是为了提升系统返回答案的召回率。最终,本文构造了一个包含103个主层次、1 750个附属层次的模板集。模板集范例如表8所示:

表8 问题模板范例

问句	类型	主层次模板	附属层次模板
新加坡的国家概况	事实类问题	[$country1] 国家 概况	[$country1] 这个 国家 怎么样
			[$country1] 国家 简介
			简单 介绍 [$country1] 大致 情况
			[$country1] 是 个 什么样 国家
			告诉 我 [$country1] 国家 情况

(2)问题模板相似度计算。问题模板相似度计算是将用户输入的自然语言问句抽象化和自动分类后,计算处理好的用户问句与模板集内模板的相似度。对于相似度计算方法,首先采用的是Word2Vec和TF-IDF相结合的方法来将问句转换成向量[28],然后利用余弦相似度(Cosine)算法[29]计算用户问句向量与模板向量的相似度,经过多次试验,笔者认为问句与模板相似度值大于0.75时,该模板选定为用户提问问题模板,当同时出现多套模板与问句相似度值大于0.75时,取相似度值最大的模板作为问题模板。

4.3.4 答案查询

程序得到问题模板后,利用问题模板对应的Cypher语句,结合识别出的实体和关系,在图数据库中查询答案,并返回给用户。用于查询具有特定关系的相关实体的Cypher模板如下:Macth(a)-[:RelationName]-(b) where b.name = 'EntityName' return a.name。其中,EntityName和RelationName用4.3.1在问句预处理中识别出的实体名和对应的关系替换。例如:对于问题"投资新加坡的国内企业有哪些?",在问题预处理后首先识别出实体名"新加坡",然后匹配模板得到该问题对应的关系为投资(investment),然后将实体名和关系嵌套入Cypher语句,查询得到答案。具体范例如表9所示:

表9 答案查询范例

问句	Cypher语言	查询结果
投资新加坡的国内企业有哪些?	Match(n1:Company)-[:investment]-(n2:Country) where n2.cname = '新加坡' return n1.ename	"浙江物产国际贸易有限公司""浙江富冶集团有限公司""南山集团有限公司"……

5 实验与结果分析

为了测试"一带一路"投资问答系统的准确性,笔者依据4.3.2问题分类,设计了6类每类30条共180条与"一带一路"投资相关的问题,对系统返回的答案

进行测评，以验证问答系统的性能。答案正确率由得到正确答案的测试问句数量与总测试问句数量的比值计算得出，公式如下：

$$\gamma=\frac{a}{c}\times 100\% \qquad 公式(1)$$

式中 γ 为答案正确率，α 为得到正确答案的测试问句数量，c 为总测试问句数量。

具体系统运行过程与实验结果详见图 5 和表 10 所示：



图 5 “一带一路”投资问答系统运行过程示例

表 10 实验结果

问题类型	测试问句数/条	答案正确数/条	正确率/%
事实类问题	30	25	83.3
方法类问题	30	27	90.0
列表类问题	30	23	76.7
计数型问题	30	24	80.0
判断类问题	30	26	86.7
其他问题	30	21	70.0

从实验结果中可以看到，系统平均回答准确率为 81.1%，绝大多数问题可以被系统正确理解并提供准确答案。尽管有些问题使用了与系统模板不一致的表述，但是由于本系统模板具有冗余性，在一定程度上提高了答案的召回率。从各类问题的回答准确率来看，准确率最高的为方法类问题，最低的为其他类问题，剩下几类问题的准确率介于方法类问题和其他类问题之间。对返回错误答案的问题进行分析，发现本系统的语义理解功能还有待进一步提升，例如：“有多少中国企业在新加坡投资？”和“在新加坡投资的中国企业有多少？”这两个问题表达的是一样的意思，但是系统在分析这两条问题时却分不清是“在中国投资的新加坡企业”还是“在新加坡投资的中国企业”，实体虽然抽取正确，但实体在语句内出现的顺序一旦颠倒，系统往往就会返回错误答案。另外，随着问题模板的不断增多，在文本分类过程中，事实类问题和其他类问题易出

现混淆,也会造成答案匹配错误。在后续研究中,笔者将考虑进一步优化系统的各部分模块提高系统的准确率,包括:①加入依存句法分析技术,深度学习技术提升系统对问题的理解能力;②进一步加大对系统对问题的覆盖率;③对知识图谱进一步完善和扩充。

在本系统的测试过程中,笔者还发现,单纯的以文本形式来回答用户提出的问题,效果并不理想,如能在系统中加入图片、视频、音频等多媒体文件可能更便于用户对答案的理解。同时,限定域问答系统具有专业化高,系统性强的特点,一般用户在使用时往往不知道系统的领域范围边界在哪,往往提出一些与系统无关的问题。因此,如何将开放域问答系统覆盖范围广、回答方式灵活等特点融入到限定域问答系统中,也是未来工作需要突破的方面。

6 结语

大数据、云计算、人工智能等技术的不断成熟和深入应用,改变了传统的信息服务方式,以新一代信息技术为支撑的信息服务具有交互更灵活、响应更快速、内容更丰富、服务移动化等特点,一方面给人们带来更多便利,另一方面也节省了大量的人力成本。因此,本文研究结果是智能信息服务的初步探索,也是践行智慧化信息服务的有益尝试,具有重要的现实意义和实践价值。笔者基于知识图谱技术,构建了"一带一路"投资问答系统。首先,在领域专家指导下根据现有公开数据资源,如《"一带一路"沿线国家投资指南》《中国对外投资企业名录》、百度知道和知乎问答等数据建立了"一带一路"投资知识图谱,在此基础上,实现了问答系统的各部分功能,包括问题预处理、问题分类、问题模板匹配和答案查询。实验表明,该系统能有效回答"一带一路"投资相关问题。下一步的工作包括进一步提高系统的语义理解功能、扩展系统可回答问题的覆盖范围、增强答案的表现能力。

参考文献:

[1] 邢厚媛."一带一路"战略下的投资促进研究[R].北京:商务部投资促进事务局,2017:2.

[2] 王雨萧,于佳欣. 我国对"一带一路"沿线国家直接投资超 900 亿美元[EB/OL].[2019-04-19].https://www.yidaiyilu.gov.cn/xwzx/gnxw/86349.htm.

[3] 郭天翼,彭敏,伊穆兰,等.自然语言处理领域中的自动问答研究进展[J].武汉大学学报(理学版),2019,65(5):417-426.

[4] 黄恒琪,于娟,廖晓,等.知识图谱研究综述[J].计算机系统应用,2019,28(6):1-12.

[5] BRILL E, LIN J, BANKO M, et al. Data-intensive question answering[C]//Trec. Tenth Text Retrieval conference. Gaithersburg: NIST,2001, 56: 90.

[6] 王东升,王卫民,王石,等.面向限定领域问答系统的自然语言理解方法综述[J].计算机科学,2017,44(8):1-8,41.

[7] LOPEZ V, MIRIAM F, MOTTA E, et al. Poweraqua: supporting users in querying and exploring the semantic web[J]. Semantic web, 2011, 3(3):249-265.

[8] BORIS K, GREGORY M, GARY B, et al. The start natural language question answering system[EB/OL].[2019-09-10].http://start.csail.mit.edu.

[9] ZHENG Z. Answerbus question answering system[C]//Proceedings of the second international conference on human language technology research. San Francisco: Morgan Kaufmann Publishers Inc., 2002: 399-404.

[10] FERRUCCI D, LEAVS A, BAGCHI S, et al. Watson: beyond jeopardy! [J]. Artificial intelligence, 2013, 199(200): 93-105.

[11] GREEN J R, BERT F, CHOMSKY C, et al. Baseball: an automatic question answerer[C]// Proceedings of the western joint computer conference. New York: IRE-AIEE-ACM, 1961:219-224.

[12] WOODS W A, KAPLAN R. Lunar rocks in natural English: explorations in natural language question answering[J]. Linguistic structures processing, 1977, 5(1): 521-569.

[13] WILENSKY R, CHIN DN, LURIA M, et al. The berkeley UNIX consultant project[J]. Artificial intelligence review, 2000, 14(1/2):43-88.

[14] VARGAS-VERA M, LYTRAS M D. AQUA: A closed-domain question answering system[J]. Information systems management, 2010, 27(3):217-225.

[15] ABACHA A B, ZWEIGENBAUM P. MEANS: a medical question-answering system combining NLP techniques and semantic web technologies[J]. Information processing & management, 2015, 51(5):570-594.

[16] ASIAEE A H, MINNING T, DOSHI P, et al. A framework for ontology-based question answering with application to parasite immunology[J]. Journal of biomedical semantics, 2015, 6(1): 1-31.

[17] XIE X,SONG W,LIU L,et al. Research and implementation of automatic question answering system based on ontology[C]//2015 27th chinese control and decision conference(CCDC). Piscataway: IEEE, 2015.

[18] ABDI A, IDRIS N, Ahmad Z. QAPD: an ontology-based question answering system in the physics domain[J]. Soft computing, 2018, 22(1): 213-230.

[19] AGARWAL A, SACHDEVA N, YADAV R K, et al. EDUQA: Educational domain question answering system using conceptual network mapping[C]//ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). Piscataway:IEEE, 2019: 8137-8141.

[20] 马晨浩.基于甲状腺知识图谱的自动问答系统的设计与实现[J].智能计算机与应用,2018,8(3):102-107.
[21] 曹明宇,李青青,杨志豪,等.基于知识图谱的原发性肝癌知识问答系统[J].中文信息学报,2019,33(6):88-93.
[22] 杜泽宇,杨燕,贺樑.基于中文知识图谱的电商领域问答系统[J].计算机应用与软件,2017,34(5):153-159.
[23] 陆伟,戚越,胡潇戈,等.图书馆自动问答系统的设计与实现[J].情报工程,2019,5(2):5-16.
[24] 刘峤,李杨,段宏,等.知识图谱构建技术综述[J].计算机研究与发展,2016,53(3):582-600.
[25] UNGER C, BUHMANN L, LEHMANN J, et al. Template-based question answering over RDF data[C]//Proceedings of the 21st international conference on world wide web. Lyon:ACM, 2012: 639-648.
[26] 王继民,王若佳,曾兰馨,等.1996-2015年“一带一路”沿线国家科研合作网络的演化分析[J].图书情报工作,2017,61(16):76-83.
[27] 张旭,孙玉伟,成颖.不同特征对文本聚类效果的比较研究——以新闻文本为例[J/OL].情报理论与实践:1-13.[2019-10-19]. http://kns.cnki.net/kcms/detail/11.1762.G3.20190903.1330.006.html.
[28] 唐明,朱磊,邹显春.基于Word2Vec的一种文档向量表示[J].计算机科学,2016,43(6):214-217,269.
[29] 高森.农业问答系统中问题分类和相似度计算的研究[D].合肥:中国科学技术大学,2018:32.

作者贡献说明：
陈璟浩：系统设计思路、论文撰写；
曾桢：系统实现、完善；
李纲：确定论文思路，提供方向建议。

A Question Answering System for “the Belt and Road” Investment Based on Knowledge Graph

Chen Jinghao[1] Zeng Zhen[2] Li Gang[3]
[1] School of Public Policy and Management, Guangxi University, Nanning 530004
[2] School of Information, Guizhou University of Finance and Economics, Guiyang 550025
[3] Center for the Studies of Information Resources of Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] The question answering system for “the Belt and Road” investment has important research and application significance. It can effectively integrate information from multiple sources and provide user with fast, accurate and high-quality “the Belt and Road” investment information. [**Method/process**] Firstly, the information which related to the “the Belt and Road” investment was collected, processed and integrated. and then, under the guidance of experts, the “the Belt and Road” investment knowledge graph was constructed. On this basis, the functions of each part of question answering system were realized, including: question preprocessing, question classification, question template matching and answer query. [**Result/conclusion**] The result shows that, this system can effectively answer questions about “the Belt and Road” investment.

Keywords: question answering system knowledge graph the Belt and Road system construction